An Introduction to Matched Pairs Designs

Wang, W. and Liu, X.

February 25, 2013
Outline

1. Related Background
   - Matching
   - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
Outline

1. Related Background
   - Matching
     - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
What is Matching?

- Matching is a strategy to remove bias in the comparison of groups by ensuring equality of distributions of the matching covariates employed.
What is Matching?

- Matching is a strategy to remove bias in the comparison of groups by ensuring equality of distributions of the matching covariates employed.

- It changes the sample frame in the analysis from the individual in an unmatched study to the matched set in matched study.
What is Matching?

- Matching is a strategy to remove bias in the comparison of groups by ensuring equality of distributions of the matching covariates employed.

- It changes the sample frame in the analysis from the individual in an unmatched study to the matched set in matched study.

- Matching is a common way to control for the effects of a covariate, such as frequency/within-class matching. A common approach to matching is to construct matched pairs.
Outline

1. Related Background
   - Matching
   - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
Matched Pairs Designs in Experiment

- It is a special case of a randomized block design.
Matched Pairs Designs in Experiment

- It is a special case of a randomized block design.
- It is used when the experiment has only two treatment conditions.
Matched Pairs Designs in Experiment

- It is a special case of a randomized block design.
- It is used when the experiment has only two treatment conditions.
- Subjects can be grouped into pairs, based on some blocking variable.
Matched Pairs Designs in Experiment

- It is a special case of a randomized block design.
- It is used when the experiment has only two treatment conditions.
- Subjects can be grouped into pairs, based on some blocking variable.
- In each pair, subjects are randomly assigned to different treatments.
Why Does Matching Reduce Bias?

- The matched pairs design is an improvement over a completely randomized design.
Why Does Matching Reduce Bias?

- The matched pairs design is an improvement over a completely randomized design.

- Like the completely randomized design, the matched pairs design uses randomization to control for confounding.
Why Does Matching Reduce Bias?

- The matched pairs design is an improvement over a completely randomized design.

- Like the completely randomized design, the matched pairs design uses randomization to control for confounding.

- But not as the other design, the matched pairs design explicitly controls for two potential lurking variables.
Why Does Matching Reduce Bias?

- The matched pairs design is an improvement over a completely randomized design.

- Like the completely randomized design, the matched pairs design uses randomization to control for confounding.

- But not as the other design, the matched pairs design explicitly controls for two potential lurking variables.

- That means matched pair designs can eliminate the bias!
Two typical Matched Pairs Designs

- Compare two treatments on one experimental unit. The block consists of the one experimental unit.
Two typical Matched Pairs Designs

- Compare two treatments on one experimental unit. The block consists of the one experimental unit.

- Compare two treatments on two very similar experimental units. The block consists of the two experimental units.
Two typical Matched Pairs Designs

- Compare two treatments on one experimental unit. The block consists of the one experimental unit.

- Compare two treatments on two very similar experimental units. The block consists of the two experimental units.

- We usually use the second form.
Outline

1. Related Background
   - Matching
   - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
Theory of Matched Study

A $2 \times 2$ frequency table:

The likelihood of the sample is

$$P(e, f, g \mid N) = \frac{N!}{e!f!g!h!} \pi_{11}^e \pi_{12}^f \pi_{21}^g \pi_{22}^h.$$
What We Care About the Study

- To test the null hypothesis of marginal homogeneity under
  \[ H_0 : \pi_{\bullet 1} = \pi_{1 \bullet} , \]
  which is equivalent to test the null of symmetry
  \[ H_0 : \pi_{21} = \pi_{12} . \]
What We Care About the Study

- To test the null hypothesis of marginal homogeneity under
  
  \[ H_0 : \pi_{11} = \pi_{1\cdot} , \]

  which is equivalent to test the null of symmetry

  \[ H_0 : \pi_{21} = \pi_{12} . \]

- To test the hypothesis of no association between exposure and disease, where the large sample McNemar test can be employed.
Outline

1. Related Background
   - Matching
   - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
Association Test: Marginal Homogeneity Test

Marginal homogeneity test under matching

\[ H_0 : P(D|E) = P_m(D|\bar{E}), \text{ or } \pi_{1\bullet} = \pi_{\bullet 1}, \]

and this is equivalent to

\[ H_0 : \pi_{12} = \pi_{21}, \]

where

\[
\pi_{1\bullet} = \int_z P(D|E, z)f_E(z)dz = P(D|E)
\]

\[
\pi_{\bullet 1} = \int_z P(D|\bar{E}, z)f_E(z)dz = P_m(D|\bar{E})
\]
Outline

1. Related Background
   - Matching
   - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
McNemar’s Large Sample Test

To test the null, the z-statistic is used:

$$Z_M = \frac{\pi_{12} - \pi_{21}}{\sqrt{(\pi_{12} + \pi_{21})/N}} = \frac{f - g}{\sqrt{f + g}},$$

which is asymptotically normally distributed under $H_0$. For a two-sided test, the equivalent chi-square statistic can be used:

$$\chi^2_M = \frac{(f - g)^2}{f + g},$$

which is asymptotically distributed as chi-square on 1 df. This is McNemar’s test (McNemar, 1947).
Outline

1. Related Background
   - Matching
   - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
Logit Model for Matched Pairs

The model is now:

\[
\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \alpha_i + \beta x_{ij}, \quad i = 1, \ldots, N; \quad j = 1, 2,
\]

where

\[
\begin{array}{c|c|c}
E (x_{i2} = 0) & D (y_{i1} = 1) & \bar{D} (y_{i2} = 0) \\
\hline
E (x_{i1} = 1) & D (y_{i1} = 1) & e = n_{11} \quad f = n_{12} \\
\hline
\bar{D} (y_{i1} = 0) & g = n_{21} \quad h = n_{22}
\end{array}
\]

\[
\pi_{i1} = P (y_{i1} = 1 \mid x_{i1} = 1) = P_i (D \mid E) = P (D \mid E, z_i)
\]

\[
\pi_{i2} = P (y_{i2} = 1 \mid x_{i2} = 0) = P_i (D \mid \bar{E}) = P (D \mid \bar{E}, z_i).
\]
Likelihoods of Logit

\[ L_i(\alpha_i, \beta) \propto \left( \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \right)^{y_{i1}} \left( \frac{1}{1 + e^{\alpha_i + \beta}} \right)^{1-y_{i1}} \left( \frac{e^{\alpha_i}}{1 + e^{\alpha_i}} \right)^{y_{i2}} \left( \frac{1}{1 + e^{\alpha_i}} \right)^{1-y_{i2}} \]

The corresponding log likelihood is

\[ \ell = \sum_i \alpha_i (y_{i1} + y_{i2}) + \sum_i y_{i1} \beta - \sum_i \log (1 + e^{\alpha_i}) - \sum_i \log (1 + e^{\alpha_i + \beta}) \]
1:M Design Logit Model

Suppose we have $n$ matched sets and that we take $i = 0$ to represent the case and $i = 1, \ldots, M$ to represent the controls. The logistic regression model is the following:

$$\text{logit}(p_j(x_{ij})) = \alpha_j + \beta^T x_{ij},$$

where $\alpha_j$ models the effect of the confounding variables in the $j$-th matched set. The conditional probability of the observed outcome, or that subject $i = 0$ is the case and the rest are controls is

$$L_j(\beta) = \frac{\exp(\beta^T x_{0j})}{\sum_{i=0}^{M} \exp(\beta^T x_{ij})}$$
Likelihood for 1:M Matching

The likelihood for the model is

$$L(\beta) = \prod_{j=1}^{n} L_j(\beta) = \prod_{j=1}^{n} \left\{ 1 + \sum_{i=1}^{M} \exp[\beta^T(x_{ij} - x_{0j})] \right\}^{-1}$$

Based on the conditional likelihood, we can make inference for the model using the information matrix.

Note that the log Odds Ratio is $\beta$:

$$\log \frac{\pi_i}{1 - \pi_i} - \log \frac{\pi_{i2}}{1 - \pi_{i2}} = \alpha_i + \beta x_{i1} - (\alpha_i + \beta x_{i2}) = \beta$$
Outline

1. Related Background
   - Matching
   - Matched Pairs Designs

2. Theoretical Results
   - Theory of Matched Study
   - Marginal Homogeneity Test
   - McNemar’s Test
   - Logit Model

3. Implementation
   - X-rays and Leukemia Example

4. Limitation
X-rays and Leukemia Data

Look at the data:

```r
> head(amlxray)

ID  disease Sex downs age Mray MupRay MlowRay Fray Cray CnRay
1  7004   1  F    no  0   no   no   no  no  no  no  1
2  7004   0  F    no  0   no   no   no  no  no  no  1
3  7006   1  M    no  6   no   no   no no  yes 3
4  7006   0  M    no  6   no   no   no no  yes 2
5  7009   1  F    no  8   no   no   no no  no  no  1
6  7009   0  F    no  8   no   no   no no  no  no  1
```
Data Manipulation

For variable *Down syndrome*, there are only seven subjects which are all cases without control. This coefficient is infinite if the variable is included. So we remove these cases for simplicity here.

```
> amlxray[amlxray$downs=="yes",1:4]

    ID disease Sex downs
   7  7010     1   M    yes
  17  7018     1   F    yes
  78  7066     1   F    yes
  88  7077     1   M    yes
 173  7146     1   F    yes
 196  7176     1   F    yes
 210  7189     1   F    yes
```
Initial Model

Mray, MupRay and MlowRay can be correlated so only Mray is picked up. The matched sets must be designated by the strata function.

```
> cmod <- clogit(disease ~ Sex+Mray+Fray+CnRay+strata(ID), ramlxray)
> summary(cmod)

Call:
  coxph(formula = Surv(rep(1, 224L), disease) ~ Sex + Mray + Fray +
     CnRay + strata(ID), data = ramlxray, method = "exact")

  n= 224, number of events= 104

          coef exp(coef) se(coef)      z  Pr(>|z|)
SexM   0.1563  1.1691   0.3861  0.405  0.68566
Mrayyes 0.2276  1.2556   0.5821  0.391  0.69573
Frayyes 0.6933  2.0003   0.3512  1.974  0.04839 *
CnRay.L 1.9408  6.9641   0.6207  3.127  0.00177 **
CnRay.Q -0.2480  0.7803   0.5819 -0.426  0.66993
CnRay.C -0.5801  0.5599   0.5906 -0.982  0.32598
```
Final Model

We pick out the significant factors and fit the model again with only linear CnRay. This time we move out Frayyes.

```r
> cmodr <- clogit(disease ~ Fray+unclass(CnRay)+strata(ID),ramlrxray)
> summary(cmodr)

Call:
coxph(formula = Surv(rep(1, 224L), disease) ~ Fray + unclass(CnRay) +
strata(ID), data = ramlrxray, method = "exact")

n= 224, number of events= 104

coef exp(coef) se(coef)     z  Pr(>|z|)    
Frayyes 0.6704  1.9550  0.3441 1.948 0.051394 .
unclass(CnRay) 0.8145  2.2580  0.2368 3.439 0.000584 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1

exp(coef) exp(-coef) lower .05 upper .05
Frayyes  1.9550  0.5115  0.996  3.838
unclass(CnRay)  2.2580  0.4429  1.419  3.592

Rsquare= 0.084  (max possible= 0.499 )
Likelihood ratio test= 19.55  on 2 df,  p=5.686e-05
Wald test  = 14.12  on 2 df,  p=0.000859
Score (logrank) test = 17.56  on 2 df,  p=0.0001539
```
Ignoring the Matching Structure

If we ignore the matching structure, we may get an incorrect analysis, which is biased.

```r
> gmod <- glm(disease ~ Fray+unclass(CnRay),family=binomial, ramlxray)
> summary(gmod)

Call:
  glm(formula = disease ~ Fray + unclass(CnRay), family = binomial,
      data = ramlxray)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.950   -0.950   -0.950    1.204    1.423

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1623    0.3011  -3.861  0.000113 ***
Frayyes      0.5004    0.3078   1.626  0.104046
unclass(CnRay)  0.6005    0.1774   3.385  0.000711 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 309.39  on 223  degrees of freedom
Residual deviance: 293.26  on 221  degrees of freedom
```
Limitation

- The more confounding variables one specifies, the more difficult to match.
The more confounding variables one specifies, the more difficult to match.

It loosens the matching requirements.
Limitation

- The more confounding variables one specifies, the more difficult to match.
- It loosens the matching requirements.
- It loses the possibility of discovering the effects of the variables used to determine the matches.


Thank you!